INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

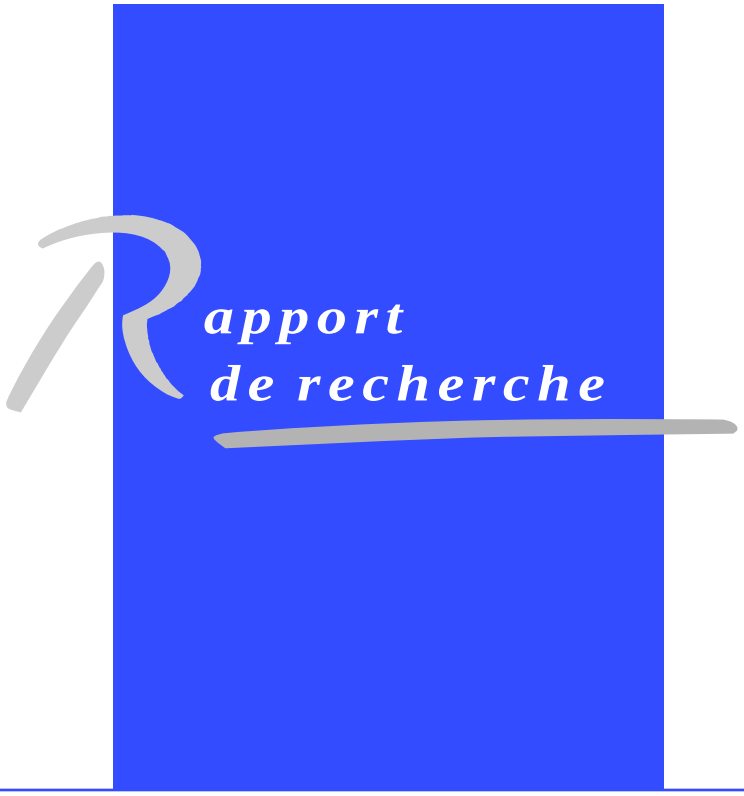
Analysis of a Yield Management Model for On Demand Computing Centers

Parijat Dube , Yezekael Hayel , Laura Wynter

N°5092

Janvier 2004

THÈME 2



ISSN 0249-6399 ISRN INRIA/RR--5092--FR+ENG

Analysis of a Yield Management Model for On Demand Computing Centers

Parijat Dube *, Yezekael Hayel †, Laura Wynter ‡

Thème 2 — Génie logiciel
et calcul symbolique
Projet Armor

Rapport de recherche n ° 5092 — Janvier 2004 — 32 pages

Abstract: The concept of yield management for IT infrastructures, and in particular for *On Demand* IT utilities was recently introduced in [13]. IT On Demand is a business model touted recently by IBM and other large IT companies as the future of computing. In [13], it was demonstrated that this paradigm is appropriate for, and can benefit greatly from, use of a convenient yield management approach, and an optimization model and theoretical justification of it were presented. The present paper provides a detailed analysis of that model, both in simplified cases where an analytical analysis is possible, and numerically on larger problem instances, and confirms the significant revenue benefit that can accrue through use of yield management in the IT On Demand context.

Key-words: Yield management, optimization, discrete choice model

(Résumé : tsvp)

This work is part of a contract between INRIA and IBM T.J. Watson Research Center.

* pdube@us.ibm.com
† yezekael.hayel@irisa.fr
‡ lwynter@us.ibm.com

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : 02 99 84 71 00 - International : +33 2 99 84 71 00
Télécopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

Etude d'un modèle de centre de calcul *à la demande*

Résumé : L'utilisation du réseau Internet et des nouvelles technologies de l'information peuvent servir au développement d'un nouveau type de service : le centre de calcul à distance. Ces nouveaux modèles de business *à la demande* sont perçus par les grandes compagnies de l'information telle IBM, comme le futur moyen de calcul à distance. Dans [13], il a été montré que les modèles de tarification basés sur l'optimisation du rendement sont appropriés. Ce papier présente une étude analytique détaillée de ce modèle dans des cas particuliers simples et numérique dans les cas plus généraux. Ces analyses confirment le bénéfice significatif en terme de revenu avec l'utilisation de techniques d'optimisation du rendement dans un contexte de services d'infrastructures *à la demande*.

Mots-clé : Optimisation, modèles de choix discrets, tarification

1 Introduction

The traditional context of IT provisioning for firms involves making periodic plans for acquisition, based on historic demand patterns combined with growth forecasts. This means that changes in IT needs are responded to with some delay, namely weeks or possibly months down the road. In addition, traditional IT provisioning, and re-sizing, requires peripheral changes in a company's profile that may cost well more than the acquisition itself, as space and human resources must be allocated to the acquisition. Analogously, a decrease in IT requirements, due to downsizing or loss of potential in some service area, is more difficult to address; due to hardware depreciation and ongoing labor contracts, downsizing IT infrastructure to reduce costs in areas or times of slow growth is not often a possibility available to firms.

In response to these business issues and based upon advances in data communication and transfer, IBM and other IT firms have begun advocating a fundamental shift in the business model underlying computing resources. The new business model, referred to as IT *On Demand*, is not actually new in the most fundamental sense; pay-for-what-you-use is a paradigm found in many sectors. However, its application to the IT and utility provisioning sector is new, and brings with it a number of challenges for the provider of an On Demand utility service, including how to jointly price and manage such a system. In particular, IT On Demand means that firms can out-source much of the IT work that is required by the firm: from servers and software, to maintenance and upgrades. The users of the firm's system do not perceive the difference; jobs and software are run remotely; web sites are stored remotely, etc. but the change is seamless from the point of view of the users. The firm then pays for IT as a service, where the price paid depends upon the usage level. In this respect, when the usage needs are high, no new acquisition need be made; it is a matter of paying per use for the extra usage consumed.

The difficulty in pricing the resource comes from the fact that jobs may arrive at will, and need not be reserved in advance. When a job arrives, the customer sending the job expects service equal to, or better than, that which they received from their proprietary, in-house systems. However, contrary to the in-house system, in an On Demand utility, many customers may expect this service on the same, shared infrastructure, at the same time. In their own, in-house infrastructure, they were assured that their needs would be met. In an On Demand utility, where the system is shared across a wide range of customers, it is no longer a trivial exercise to guarantee a given level of service to everyone without possibly an a priori reservation of resources. Consider the case where several large customers expecting a high quality of service arrive at overlapping intervals. Managing the system thus becomes a complex task. In addition, and equally important, devising service offerings and pricing strategies, before customers join the system, that take into account the management of the system, is complex but essential.

In this paper, we analyze the model introduced in [13], both analytically and numerically. The analytical study is carried out on simplified versions of the model and when the number of variables is small; as such it provides a bound on what can be said about the full-scale

model. The numerical study then illustrates the benefits that the approach and our model can provide and confirms the tractability of the yield management paradigm to the IT On Demand context.

The structure of the paper is as follows. In Section 2, we present a review of related literature. Then, in Section 3, we recall the model of [13] and necessary notation. In Section 4, we present the analytical study of the model in several simplified settings. Section 5 contains the numerical study of the model. Finally, we conclude in Section 6 with a number of valuable directions for further research in the area.

2 Literature review

Models for determining yield management-type pricing and seat booking limits for airlines have evolved from the original single-leg capacity rationing problem to more complex models which, for example, consider networks of flights and connections, as well as other sectors characterized by variable-duration activities, such as hotels. In the former setting, the number of seats to offer at each pre-determined price class, on a single, isolated, flight is determined. (see, e.g., Belobaba [1] or Littlewood [8]) The latter setting is more complex due to the size of the problem under study and simplified versions of it, such as the so-called nesting approach which decomposes the problem into single-leg subproblems, are generally presented as more practicable than solving the complete problem. (See the survey of McGill and van Ryzin [11], and references therein).

These models often make simplifying assumptions that the demand model underlying airline passenger choices is exogenously given and fixed, or other times assume a news-vendor-type formulation in which demands are treated like continuous fluids that can be partially accepted ([11], for example, uses the latter formulation). A control-theoretic formulation of Kleywegt [7] falls into this latter category as well.

In many other fields of management science, however, the assumption that demand for a good or service depends upon the characteristics of that good or service which are under study, is taken for granted. Indeed, demand is modelled as an endogenous variable, or process. In the yield management literature, this connection has been taken into account heuristically at times through the so-called buy-up probabilities, which give, for each pair of fares, the probability that a user "buys-up" to the higher-cost fare when the lower fare is sold out. However, as pointed out by [12], this pair-wise approach is not realistic when considering more than two or three fare classes. The authors in [12] formulate a simplified (single-leg) model for the airline yield management problem when the choice probabilities of customers are described by the multinomial logit model. The authors also discuss model estimation in that reference.

While the literature on airline yield management is clearly of great relevance to our problem of yield management in On Demand solution centers, there are notable differences which lead to significant higher complexity in our setting. First and foremost, the good

or service under consideration in IT on demand does not have a fixed duration nor does it occupy a pre-determined percentage of the resource capacity. That is, an airline seat is occupied precisely for the duration of the flight, and the number of seats to sell on any flight is known in advance. On the other hand, in On Demand IT utilities, the duration of a job depends upon the type of server upon which it is run, and the number of servers, if it is parallelizable; further, the number of servers it requires depends upon the type of servers that are used. In other words both the capacity needed and the time taken by a job are not simple, exogenous parameters in the compute On Demand yield management problem. Some features of this time variability can be observed in other sectors, such as hotellerie, restaurant yield management, and even golf course yield management (see, for example, [5] and [6] and other references by those authors). Nonetheless, the capacity and percentage of capacity occupied in these latter examples are still fixed and exogenous, as opposed to the setting with which we are faced.

Work on the pricing of information services, such as the pricing strategies of internet service providers (ISPs) has traditionally considered some of these issues of job duration and capacity occupation through queueing formulae. The literature on that and related areas is quite vast and a thorough survey of it is not the focus of our work here, but a few relevant references are [10],[4],[9]. The difference between the decisions optimized in those and related work is the degree of segmentation. In the internet pricing world, a single price per type of service is proposed. It is sometimes the case that multiple qualities of service (QoS) are discussed, but in that case, each QoS level has associated with it a single, fixed price. The number of such price levels is generally limited to three, for example, gold, silver, and bronze-level service. The yield management strategy takes customer segmentation to a much finer level, and does so through the incorporation of demand models.

3 The model

The optimal yield management reservation system parameters for an *On-Demand* IT computing center can be obtained through the following model.

We assume that the On Demand utility is composed of a pool of nodes (processing units) and a pool of storage space to allocate to different fee classes. Let there be Q groups of nodes with $\mathcal{N}_q$, $q = 1, \ldots, Q$ being the number of nodes in the qth group. The nodes in a class are then assumed to be homogeneous, have the same speed, etc. The storage pool is also of finite size, given by some constant, $\mathcal{S}$.

The optimization problem that we will need to solve is then the following: At each time epoch $i, i = 1 \ldots N$ we would like to reserve the available resources (nodes and external storage) for the different fee classes. The resources should be allocated so as to maximize expected provider profits, that is, expected revenue less expected costs, where expected provider revenue is related to the distributions of different customer arrival types, their

preferences (in terms of service/price tradeoffs) as well as their service requirements, and to the number of nodes assigned to each fee class, on each server type.

Fee classes are defined much as in the usual yield management literature. For an identical resource several different prices may co-exist. Each fee class then has a maximal number of users, and once that number is reached within the time period for that fee class, new requests are offered only the next higher level fee for that resource.

Resources are also defined in a broad way. While a server and storage are clearly aspects of the resource, the service-level (SLA) parameters are as well, such as availability, advance notice, penalties in case of non-satisfaction of service level by the provider, etc. The broad scope of the *resource* in this manner allows the price differentiation to become still finer-grained; that is, for an identical server/storage combination, different SL offerings create new sets of fee classes.

We next define the basic notation and assumptions.

3.1 Notation and Assumptions

Let us define

i = time epoch index ($i = 1, \ldots, N$)
k = fee class index ($k = 1, \ldots, K$)
q = machine (or, node) type index ($q = 1, \ldots, Q$)
c = customer class, indicates type of workload, $c = 1, \ldots, C$
n_{ikq} = decision variable: # of nodes of type q allocated to fee class k at time i
s_{ik} = decision variable: amount of storage assigned to fee class k at time i
Q_{ik} = set of q's at time i for the kth class for which $n_{ikq} > 0$
r_{ikq} = per unit profit=revenue-cost of a node of type q for fee class k at time i
p_{ik} = per unit profit of storage for fee class k at time i
N_q = total number of nodes of type q available
$\mathcal{S}$ = total amount of storage available
$\mathcal{S}_i$ = total amount of storage available at time i, a random variable with a discrete distribution
v_q = processor speed of machine of type q
β_c = fraction of work that is not parallelizable for a job of type c
$P_{ik}(t)$ = probability that a customer with workload t arriving at time i offered fee class k on machine type q accepts the offer (discrete choice)

Γ_c = probability that an arriving job is of type c
W_c = random workload of a job of type c (a discrete RV)
S_i = random storage requirement of a job entering at time i

Assumption 1 *[Heterogeneous resources] We shall assume in this work that the pool of servers, or nodes, is heterogeneous, but that the distinguishing feature of the different server types is their processing speeds. We shall neglect other differences that may be present when dealing with heterogeneous resource pools.*

Assumption 2 *[Distributed computing] We shall assume that jobs can be processed in parallel, up to a degree specified by a* serial-part *parameter, β_c, given once the type of the job c is given. For example, for $c = 1$ we have some β_1, such as $\beta_1 = 1$, where 1 means that the serial part of jobs of type 1 is 100% of their workload. Similarly, we may set $\beta_2 = 0.75$ so that 25% of type 2 jobs are parallelizable. This definition allows for computation of the processing time T as a function of the configuration given to the user for that job. A precise formula is provided in Section 3.3.*

3.2 Optimization formulation

The goal of the optimization model is to determine the optimal reservation of resources across server types, time epochs, and fee classes, so as to maximize expected revenue. This can be formalized by the following optimization problem:

$$\max_{n_{ikq}, s_{ik}} E\left[\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{k=1}^{K} T_i(W_i, n_{ikq}, c)\left(\sum_{q=1}^{Q} r_{ikq} n_{ikq} + p_{ik} s_{ik}\right) P_k(W_i, n, s)\Gamma_c\right],$$

where T_i is the (possibly random) sojourn time of a job in the system at time i, and depends upon the workload, or size, of the job, W_i, the number of slots allocated n_{ikq}, and the type of job, c . The choice probability of a user accepting a slot of segment-type k is given by $P_k(W_i, n, s)$, and the probability of an arrival of job class c is given by Γ_c. The decision variables are n_{ikq}, the number of processing slots to reserve at time i, in price segment k, on machine type q, and the amount of storage to reserve at time i, at fee class s_{ik}. The parameters r_{ikq} and p_{ik} are the possible prices at which those resources can be reserved. By enumerating a wide range of such prices, the optimization model works by identifying those price segments which are most profitable to offer, given the characteristics of the available demand and resource levels.

Observe that once a customer joins class k at time i then it takes n_{ikq} nodes and these nodes are only released when the customer finishes it job and leaves the system. Thus these nodes will be released at time $i + T_i(W_i, n_{ikq}, c)$ and from time i to $i + T_i(W_i, n_{ikq}, c)$ these resources cannot be allocated to other classes.

We assume that the total amount of storage space $\mathcal{S}$ is finite, as is the total number of nodes. Furthermore, the storage allocated must be at least as large as the expected amount requested by a time i job. The constraints on those resources are provided below.

$$N_q - \sum_{k,z \leq i, z+T_i(W_i,n_{ikq},c)>i} n_{zkq} \geq 0, \forall i,q \quad (1)$$

$$\mathcal{S} - \sum_{k,z \leq i, z+T_i(W_i,n_{ikq},c)>i} s_{zk} \geq 0, \forall i \quad (2)$$

$$\sum_{k=1}^{K} s_{ik} - E[S_i] \geq 0, \forall i \quad (3)$$

$$n_{ikq} \geq 0, \forall i,k,q \quad (4)$$

Alternatively, one can assume that the resource limits are *soft constraints* and include the possibility to surpass those limits, at a cost associated with having to make use of remote resources or to pay a penalty to the customers. Note that since the total allocated CPU and memory at any epoch cannot exceed the authorized amount, we are not allowing for *overbooking* in this formulation.

3.3 Expression for the sojourn time, $T_i(W_i, n_{ikq}, c)$

There are numerous ways to compute the sojourn time of a job on one or more servers, depending on the data that the user provides about the job. We consider three ways in which this computation can be done.

1. In the simplest case, the user may provide an estimated duration of the job for different server speeds.
2. Often, however, the user may not have such information available. Instead he may specify his job in terms of its *type,* (CPU intensive, memory intensive, storage intensive etc.) and workload size, or time on a single processor of given speed, and a set of formulæ can be used to determine the sojourn time on the configuration presented by the provider, which may involve a processor of a different speed than the one given by the user as a reference, and/or multiple processors, in which case an Amdahl-type law can be used to provide an estimate of the combined serial-parallel computation time.
3. Finally, a more complex, fine-grained analysis could be performed in which sojourn times are predicted using queuing models and simulation.

We make use of methods 1 and 2. Indeed, the advantage of linking sojourn time directly to the decision variables, as in method 2, is countered by the fact that the resulting overall problem becomes highly non-convex, even when simplifying assumptions are made on the other distributions/

3.3.1 Independent, fixed or externally-provided sojourn times

In this case, we make use of an externally-provided (e.g., from the user making the request) estimation of the average sojourn time of the job in question; i.e., the sojourn time is a *constant* for each three-tuple of job entry time—fee class–server type(s), (i, k, q). Notably, this case is useful as it simplifies the mathematical properties of the optimization model and provides some average-case value for the results.

3.3.2 Configuration-dependent sojourn times

We also consider a more detailed, configuration-dependent expression for the sojourn times, where the sojourn time is a function of the decision variables, that is the number of nodes allocated to the job. The formula is provided below. An analysis of the resulting properties of the optimization model is presented in a later section.

Let us consider a job of size W_i. Let β_c be the fraction of job that can only be executed serially for a job of type c, and $1-\beta_c$ is then the fraction of job that can be parallelized. We thus have the following bound on the actual sojourn time of a job on one or more nodes:

$$\overline{T_i}(W_i, n_{ikq}, c) = \frac{\beta_c W_i}{\min_{q \in Q_i} v_q} + \frac{(1-\beta_c)W_i}{\sum_{q \in Q_i} v_q n_{ikq}}. \tag{5}$$

Proposition 1 *The sojourn time, $T_i(W_i, n_{ikq}, c)$ is bounded from above by $\overline{T_i}(W_i, n_{ikq}, c)$, that is $T_i(W_i, n_{ikq}, c) \leq \overline{T_i}(W_i, n_{ikq}, c)$ for every i, k, q, c.*

Proof. The expression on the right-hand-side of (5) is given by an application of Amdahl's Law to our context, where the serial processing time of the job is given by W_i/v_q for a job of workload size W_i processed on a single server of type q. Inequality (5) is indeed a bound as, for the serial job (the first term of the right-hand-side), we are assuming that it is executed on the slowest node among the assigned nodes, where Q_i is the set of q's for which $n_{ikq} > 0$. □

3.3.3 Deterministic equivalent–expected satisfaction of constraints

Observe that the node allocation constraints are stochastic, as the sojourn times, $T_i(.)$, are random. We propose to replace them by deterministic constraints, through an expected-value approach.

$$E\left[N_q - \sum_{k,z \leq i, z+T_i(W_i, n_{ikq}, c) > i} n_{zkq}\right] \geq 0, \forall i, q, \tag{6}$$

$$E\left[\mathcal{S} - \sum_{k,z \leq i, z+T_i(W_i, n_{ikq}, c) > i} s_{zk}\right] \geq 0, \forall i. \tag{7}$$

Conditioning on W_i assuming a discrete distribution on that random variable, we can write from (6) and (7):

$$\sum_{c=1}^{C}\sum_{w=0}^{\infty}\left(N_q - \sum_{k,z\leq i, z+T_i(W_i,n_{ikq},c)>i} n_{zkq}\right) P(W_i=w)\Gamma_c \geq 0, \forall i,q, \tag{8}$$

$$\sum_{c=1}^{C}\sum_{w=0}^{\infty}\sum_{s=0}^{\infty}\left(\mathcal{S} - \sum_{k,z\leq i, z+T_i(W_i,n_{ikq},c)>i} s_{zk}\right) P(W_i=w)\Gamma_c \geq 0, \forall i. \tag{9}$$

3.4 Choice function of user preferences

The probability that a user chooses one of a finite, discrete set of options is elegantly described by the logit function. Indeed, the logit function has a number of desirable properties; notably, the choice probability function is simple to express in closed form and to evaluate. Also, it subsumes the case of pure randomness and purely deterministic choice through a single parameter, referred to here as θ. Finally, it handles different types of variables; when variable values are binary, for example, it has been shown to be more robust than linear regression, and can also handle ratings and rankings, in addition to usual continuous variables. (See, for example, [2])

Let ζ_1 be the user sensitivity to quality of service (here, given by sojourn time) and ζ_2 to the total price he will pay. We assume that the disutility is linear in its parameters. The number and nature of these parameters is quite flexible; in particular, we may consider parameters describing the availability level of the service, the minimum advance-notification time, as well as characteristics about the user, such as whether it is a frequent client or not.

Here we define, for ease of presentation, a simple disutility function composed of only price and time parameters, respectively. In this case, we obtain the disutility that a user arriving at time i, assigned to class k experiences is:

$$\mathcal{U}(k) = \zeta_1 T_i(W_i, n_{ikq}, c)\left(\sum_{q=1}^{Q} r_{ikq}n_{ikq} + p_{ik}s_{ik}\right) + \zeta_2 T_i(W_i, n_{ikq}, c).$$

The constants of the disutility function may be calibrated from empirical data. The choice of a user to accept fee class k may be modelled as deterministic (and rational) or stochastic, (i.e., with some degree of irrationality). In other words, a deterministic and rational choice would be for a user to choose the $\hat{k}$ that solves $\min_{k=1..K}\mathcal{U}(k)$. However, due to misinformation, unmeasurable parameters that play a role in user choices, and/or measurement error, user choices may be best represented by stochastic models.

The choice probability P_k is therefore given by the multinomial logit function as follows:

$$P_k = \frac{e^{-\theta \mathcal{U}(k)}}{\sum_{j=1}^{K} e^{-\theta \mathcal{U}(j)}}.$$

The desired value of $\theta \in [0, \infty]$ is problem-dependent, and involves the degree of randomness that one supposes is present in the data, as well as the size of the parameters in the linear disutility function.

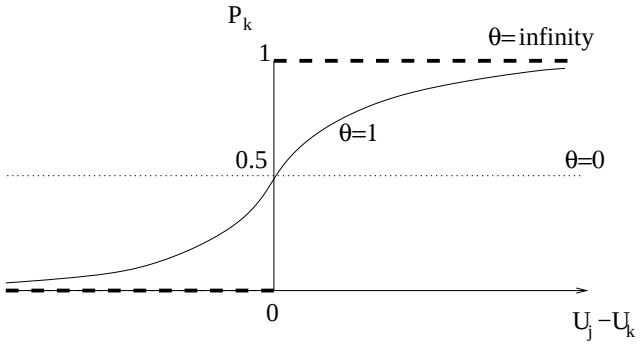


Figure 1: The logit choice probability as a function of the difference of two dis-utilities, for three values of theta.

Using the logit scaling parameter, θ, we can isolate three distinct cases, illustrated in Figure 1.

1. $\theta = 0$. In this case, the choices are purely random, as $P_k = 1/K$ for each choice $k = 1 \ldots K$. In the Figure, since K=2, the probability is $1/2$ for each choice irrespective of the utilities. Note that if this is the behavior that we wish to model, we may substitute $P_k = 1/K$ explicitly. This will be of use in determining special, linear, cases of our optimization model.

2. $\theta = \infty$. In this case, we also obtain a special case of the optimization model. By setting $\theta = \infty$, we are in effect setting $P_k = \min_{j=1 \ldots K} U_j$, which is a step function, and represents a purely deterministic choice of selecting the minimal disutility option. Note that this function is therefore non-differentiable, as opposed to the other two cases.

3. When $\theta \in]0, \infty[$, the logit function is differentiable, but nonlinear over $\Re$. It implies that there is some randomness, or error in the perception of the true costs (or dis-utilities) of each choice, hence the spread of choices about the true minimum cost choice.

3.5 Objective function with logit choice probabilities and sojourn time expression

Substituting the expression for P_k, the choice probabilities, from the definition of the multinomial logit model, we can write the complete objective function as follows:

$$\max_{n_{ikq},s_{ik}} E\left[\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{k=1}^{K} T_i(W_i,n_{ikq},c)\left(\sum_{q=1}^{Q} r_{ikq}n_{ikq}+p_{ik}s_{ik}\right)\frac{e^{-\theta\mathcal{U}(k)}}{\sum_{j=1}^{K}e^{-\theta\mathcal{U}(j)}}\Gamma_c\right].$$

Conditioning on the workload and storage requirement of the customer we get with (5):

$$\max_{n_{ikq},s_{ik}} \sum_{w=1}^{\infty}\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\frac{\beta_c w}{\min_{q\in Q_i} v_q}+\frac{(1-\beta_c)w}{\sum_{q\in Q_i} v_q n_{ikq}}\right)\left(\sum_{q=1}^{Q} r_{ikq}n_{ikq}+p_{ik}s_{ik}\right)$$
$$\times\frac{e^{-\theta\mathcal{U}(k)}}{\sum_{j=1}^{K}e^{-\theta\mathcal{U}(j)}}\Gamma_c P(W_i=w) \tag{10}$$

Further, we have that $\min_{q\in Q_i} v_q \leq \xi \equiv \min_{q\in Q} v_q$. Thus we can simply the formulation with this bound:

$$\max_{n_{ikq},s_{ik}} \sum_{w=1}^{\infty}\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\frac{\beta_c w}{\xi}+\frac{(1-\beta_c)w}{\sum_{q\in Q_i} v_q n_{ikq}}\right)\left(\sum_{q=1}^{Q} r_{ikq}n_{ikq}+p_{ik}s_{ik}\right)$$
$$\times\frac{e^{-\theta\mathcal{U}(k)}}{\sum_{j=1}^{K}e^{-\theta\mathcal{U}(j)}}\Gamma_c P(W_i=w) \tag{11}$$

subject to constraints (1)-(4), with the deterministic equivalents, (8)-(9) in place for (1)-(2), when the latter are random.

4 Analytical study of the model

In this section, we consider a simplified model in which two different prices per node are offered, i.e. $r_1 \neq r_2$. Furthermore, we shall consider two different user, or job, classes, c. We assume here that T is exogenously given, but depends on the customer class c. We shall also consider one single time epoch $i=1$, one type of processing node $Q=1$ and we focus on a model without the storage component. Under these simplifications, we shall be able to examine analytically the Hessian of the Lagrange function and solve to obtain bounds on the decision variables. The simplified problem can be expressed as:

$$\max_{n_1,n_2\geq 0} F(n_1,n_2)=\sum_{c=1}^{C}\sum_{k=1}^{2} T_c r_k n_k P_k^c \Gamma_c,$$
$$n_1+n_2=N.$$

While these results cannot be extended in general for any number of parameters, they, along with the larger-scale numerical results, provide valuable insight into the nature of the problem under study.

Two discrete choice models are considered below. In the first, the proportion of individuals choosing option i is *deterministic*, and given by the normalized ratio of the utility of choice i to the sum of all choice utilities, that is, $\forall c = 1, \ldots, C$,

$$P_k^c = \frac{1}{K-1}\left(1 - \frac{U_k^c}{\sum_{j=1}^{K} U_j^c}\right).$$

The first term normalizes the quantities P_k^c so that they sum to 1 for each customer class, c. The second term is expressed as $1-$ ratio, since the U_k^c are actually dis-utilities and hence decreasing in price and delay. With only two price segments, as we have stated would be the case throughout this section, we have, for every c,

$$P_1^c = \frac{U_2^c}{U_1^c + U_2^c} \quad \text{and} \quad P_2^c = \frac{U_1^c}{U_1^c + U_2^c},$$

where the (dis-)utility functions are:

$$U_k^c = \zeta_1 T_{kc} r_k n_k + \zeta_2 T_{kc}.$$

The parameters ζ_1 and ζ_2 are taken as constants here. Recall also that T_{kc} is a constant in this section. The utility function is thus linear in the decision variable, n_k.

The second discrete choice model considered is the multinomial logit choice function, where the probability of a user of class c choosing price k is defined by:

$$P_k^c = \frac{e^{-\theta U_k^c}}{\sum_{j=1}^{K} e^{-\theta U_j^c}}, \forall c,$$

and θ is the control parameter that determines the degree of randomness of the user choice model, where $\theta = 0$ means that the choice is purely random and does not depend upon the utilities (but rather is constant at $1/K$) and $\theta = \infty$ means that the choice is purely deterministic in that when k is the minimum disutility choice, its probability of being chosen is equal to 1, and the probability of choosing any other option, $k \neq j$ is 0. Here, as we have set the number of price segments $K = 2$, we obtain the pair of logit preference functions:

$$P_1^c = \frac{1}{1 + e^{\theta(U_1^c - U_2^c)}} \quad \text{and} \quad P_2^c = \frac{1}{1 + e^{\theta(U_2^c - U_1^c)}}.$$

4.1 Analysis using the deterministic discrete choice model

Explicitly expressing the deterministic discrete choice model into the objective function for this two-segment example, we obtain:

$$\max_{n_1,n_2\geq 0} F(n_1,n_2) = \sum_{c=1}^{C} \Gamma_c T_c \left(f(n_1,n_2) + g(n_1,n_2)\right) \qquad (12)$$
$$n_1 + n_2 \leq N$$

with

$$f(n_1,n_2) = r_1 \frac{\zeta_1 r_2 n_1 n_2 + \zeta_2 n_1}{\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2},$$

and

$$g(n_1,n_2) = r_2 \frac{\zeta_1 r_1 n_1 n_2 + \zeta_2 n_2}{\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2}.$$

As the sojourn time is decreasing with the number of nodes, the inequality constraint expressed in (12) is active in real situations. In this simplified setting, we have the following result.

Proposition 2 *The nonlinear yield management reservation problem of 12 has a unique maximum.*

Proof: The Lagrangian function is :

$$L(n_1,n_2,\mu) = \sum_{c=1}^{C} \Gamma_c T_c.(f(n_1,n_2) + g(n_1,n_2)) - \mu G(n_1,n_2),$$

with Lagrange multiplier μ, and constraint function: $G(n_1,n_2) = n_1 + n_2 - N$.

At optimality, the constraint will be active. Hence, after some manipulation of the Lagrangian, we obtain the following system:

$$\begin{cases} \frac{r_1 \sum_{c=1}^{C} \Gamma_c T_c (2\zeta_1^2 r_2^2 n_2^2 + 4\zeta_1\zeta_2 r_2 n_2 + 2\zeta_2^2)}{(\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2)^2} - \mu = 0, \\ \frac{r_2 \sum_{c=1}^{C} \Gamma_c T_c (2\zeta_1^2 r_1^2 n_1^2 + 4\zeta_1\zeta_2 r_1 n_1 + 2\zeta_2^2)}{(\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2)^2} - \mu = 0, \\ n_1 + n_2 = N, \end{cases}$$

from which we obtain:

$$\begin{cases} (\zeta_1 r_2 n_2 + \zeta_2)^2 - \left(\sqrt{\frac{\mu}{2r_1 \sum_{c=1}^{C} \Gamma_c T_c}}(\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2)\right)^2 = 0, \\ (\zeta_1 r_1 n_1 + \zeta_2)^2 - \left(\sqrt{\frac{\mu}{2r_2 \sum_{c=1}^{C} \Gamma_c T_c}}(\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2)\right)^2 = 0, \\ n_1 + n_2 = N. \end{cases}$$

Noting that each of the first two equations is a quadratic, we obtain four sets of equations. However, of those, only one has a feasible solution, which we shall show is the unique maximum, where the solution is given by:

$$(n_1^*, n_2^*) = (N\frac{\sqrt{r_2}}{\sqrt{r_1}+\sqrt{r_2}} + \mathcal{H}, N\frac{\sqrt{r_1}}{\sqrt{r_1}+\sqrt{r_2}} - \mathcal{H}),$$

with $\mathcal{H} = \frac{\zeta_2}{\zeta_1 r_2}\frac{\sqrt{r_2}}{\sqrt{r_1}+\sqrt{r_2}}(1 - \sqrt{\frac{r_2}{r_1}})$. The optimal Lagrangian multiplier, μ^*, is:

$$\mu^* = 2\sum_{c=1}^{C} \Gamma_c T_c \frac{r_1 r_2}{(\sqrt{r_1}+\sqrt{r_2})^2}.$$

Additionally, coming from n_1^* and n_2^* non-negative, we obtain two necessary conditions on the feasible range of prices, r_1 and r_2, for a solution:

$$r_1 \geq \frac{r_2}{(1 + N\frac{\zeta_1 r_2}{\zeta_2})^2}, \tag{13}$$

$$r_1 \leq r_2(1 + N\frac{\zeta_1 r_2}{\zeta_2})^2. \tag{14}$$

Thus, depending on the values of the problem constants, we can determine the range of prices for which a solution exists. Figure 2 illustrates this range for a capacity level $N = 10$, and for a range of value-of-time parameters, $\zeta = \zeta_1/\zeta_2$.

To prove that the solution we have found is indeed a maximum, we use the Hessian of the Lagrange function defined by:

$$\overline{D} = \begin{pmatrix} 0 & G_{n_1} & G_{n_2} \\ G_{n_1} & L_{n_1 n_1} & L_{n_1 n_2} \\ G_{n_2} & L_{n_2 n_1} & L_{n_2 n_2} \end{pmatrix}$$

If the determinant of the Hessian of the Lagrangian function $\overline{D}$, evaluated at the possible optimum, is positive thus the optimum point is a maximum. The second derivatives of the Lagrangian function are:

$$L_{n_1 n_1} = \frac{\partial^2 L}{\partial n_1^2}(n_1, n_2) = -\sum_{c=1}^{C} \Gamma_c T_c (2\zeta_1^2 r_1 r_2^2 n_2^2 + 4\zeta_1 \zeta_2 r_1 r_2 n_2 + 2r_1 \zeta_2^2)\frac{2\zeta_1 r_1}{(\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2)^3},$$

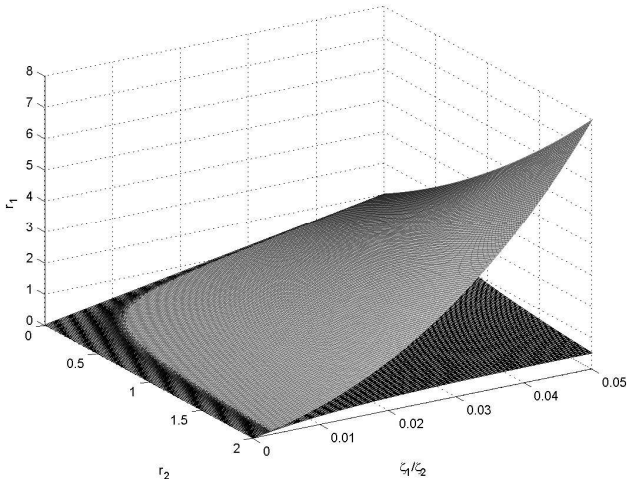


Figure 2: Existence area of the solution when $N = 10$.

$$L_{n_2n_2} = \frac{\partial^2 L}{\partial n_2^2}(n_1,n_2) = -\sum_{c=1}^{C} \Gamma_c T_c (2\zeta_1^2 r_1^2 r_2 n_1^2 + 4\zeta_1\zeta_2 r_1 r_2 n_1 + 2r_1\zeta_2^2) \frac{2\zeta_1 r_2}{(\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2)^3},$$

$$\begin{aligned} L_{n_1n_2} = \frac{\partial^2 L}{\partial n_1 n_2}(n_1,n_2) = \frac{\sum_{c=1}^{C} \Gamma_c T_c}{(\zeta_1 r_1 n_1 + \zeta_1 r_2 n_2 + 2\zeta_2)^3} \big[& 4\zeta_1^4 r_1^3 r_2^2 n_1^2 n_2 + \\ & + 4\zeta_1^4 r_1^2 r_2^3 n_1 n_2^2 + 16\zeta_1^3\zeta_2 r_1^2 r_2^2 n_1 n_2 + 8\zeta_1^3\zeta_2 r_1 r_2^3 n_2^2 + \\ & + 4\zeta^3\zeta_2 r_1^3 r_2 n_1^2 + 8\zeta_1\zeta_2^3 r_1 r_2 + 12\zeta_1^2\zeta_2^2 r_1^2 r_2 n_1 + 12\zeta_1^2\zeta_2^2 r_1 r_2^2 n_2 \big]. \end{aligned}$$

The determinant of the Hessian is:

$$|\overline{D}| = -G_{n_1}^2 L_{n_1n_1} + 2G_{n_1} G_{n_2} L_{n_1n_2} - G_{n_2}^2 L_{n_2n_2},$$

and $G_{n_1} = G_{n_2} = 1$. Thus we obtain that the determinant of $\overline{D}$ is positive for all $n_1 \geq 0$ and $n_2 \geq 0$, and we conclude that (n_1^*, n_2^*) is a maximum over the frontier. Moreover the objective function is continuous, this point is a global maximum.

4.2 Analysis using the logit choice model

When the customer choice function is given by the probabilistic logit model, we obtain the following nonlinear maximization problem

$$\begin{aligned} \max_{n_1,n_2 \geq 0} F(n_1,n_2) &= \sum_{c=1}^{C} \Gamma_c T_c \left(r_1 n_1 \frac{1}{1 + e^{\theta\zeta_1 T_c (r_1 n_1 - r_2 n_2)}} + r_2 n_2 \frac{1}{1 + e^{\theta\zeta_1 T_c (r_2 n_2 - r_1 n_1)}} \right) \\ n_1 + n_2 &= N \end{aligned} \tag{15}$$

As before, the revenue is maximum when all resources are occupied, which implies that the inequality constraint will be active at the solution. In this case, to simplify the added complexity posed by the logit model, we shall make use of the equality constraint in eliminating one of the two decision variables from the formulae, since they sum to a constant. To further simplify the logit model for the sake of analytical analysis, we assume here that the total workload is the same for the different customer classes, i.e. $T_1 = T_2 = \ldots = T_c = T$. In this case, the probability term of each workload type, Γ_c, vanishes, and we thus obtain the following objective function, where the first and second terms of (15) are labelled $f(n_1)$ and $g(n_1)$, resp:

$$F(n_1) = T(f(n_1) + g(n_1)).$$

Setting $y(n_1) = e^{\theta\zeta_1 T n_1(r_1+r_2)-\theta\zeta_1 T r_2 N}$, after some manipulation, the derivative of the objective function with respect to the single variable, n_1, is

$$F'(n_1) = T\left(\frac{r_1}{(1+y)^2}(1+y-n_1y') - y^2\frac{r_2}{(1+y)^2}(1+y^{-1}-(N-n_1)y'/y^2)\right).$$

We thus must solve the following equation:

$$y^2r_2 + y(r_2-r_1) + y'(r_1n_1 - r_2(N-n_1)) = r_1. \tag{16}$$

In addition, we have that:

$$y'(n_1) = \theta\zeta_1 T(r_1+r_2)y(n_1),$$

and thus Equation (16) becomes

$$y^2r_2 + yH(n_1) = r_1,$$

with $H(n_1) = ((r_2-r_1) + \theta\zeta_1 T(r_1+r_2)(r_1n_1 - r_2(N-n_1)))$.
We obtain:

$$y\sqrt{r_2} = \sqrt{\frac{H(n_1)^2}{4r_2} + r_1} - \frac{H(n_1)}{2\sqrt{r_2}}.$$

Thus, to find the optimal n_1^* we have to solve the following equation:

$$\begin{aligned} e^{\theta\zeta_1 T n_1(r_1+r_2)-\theta\zeta_1 T r_2 N} &= \frac{1}{\sqrt{r_2}}\sqrt{\frac{H(n_1)^2}{4r_2} + r_1} - \frac{H(n_1)}{2r_2}, \\ n_1 &= \frac{\ln(\sqrt{H(n_1)^2 + 4r_1r_2} - H(n_1)) - \ln(2r_2) + \theta\zeta_1 T r_2 N}{\theta\zeta_1 T(r_1+r_2)}, \\ &= M(n_1) \end{aligned} \tag{17}$$

The left-hand side is strictly increasing as its derivative is:

$$(e^{\theta\zeta_1 T n_1(r_1+r_2)-\theta\zeta_1 T r_2 N})' = \theta\zeta_1 T r_1 e^{\theta\zeta_1 T n_1(r_1+r_2)-\theta\zeta_1 T r_2 N} > 0,$$

and $M(n_1)$ is strictly decreasing as its derivative is:

$$(\frac{1}{\sqrt{r_2}}\sqrt{\frac{H(n_1)^2}{4r_2}+r_1}-\frac{H(n_1)}{2r_2})' = \frac{\theta\zeta_1 T(r_1+r_2)^2}{2r_2}\left(\frac{H(n_1)}{\sqrt{H(n_1)^2+4r_1r_2}}-1\right) < 0.$$

Hence, we have a necessary and sufficient condition to the existence of a valid solution of the fixed point equation.

Condition 1 *The first part of the condition is given by $M(0) \leq N(0)$, i.e.*

$$r_1 \geq e^{-\theta\zeta_1 T r_2 N}\left(r_2 - r_1 - \theta\zeta_1 T(r1+r2)Nr_2 + r_2 e^{-\theta\zeta_1 T r_2 N}\right).$$

The second part, given by $M(N) \geq N(N)$, is:

$$e^{\theta\zeta_1 T r_1 N}\left(r_2 - r_1 + \theta\zeta_1 T(r1+r2)Nr_1 + r_2 e^{\theta\zeta_1 T r_1 N}\right) - r_1 \geq 0.$$

In Figure 3, we present an example of this fixed point problem with (unique) job time, $T = 2$, logit scaling parameter, $\theta = 0.05$, value of time constant, $\zeta_1 = 1$, prices at the two price levels, $r_1 = 3$, $r_2 = 6$ and total available capacity, $N = 10$. In this particular case, the optimal solution is $n_1^* = 6.2892$ and $n_2^* = 3.7108$ slots at each of the two price levels.

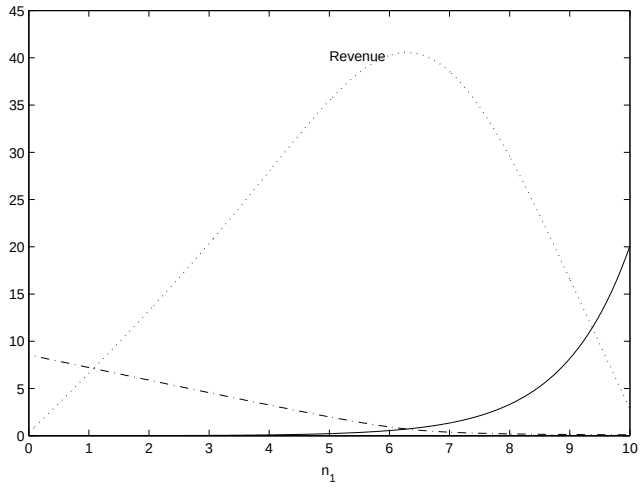


Figure 3: Solution with 2 classes and logit discrete choice model.

It is furthermore possible to express analytically the optimal solution (n_1^*, n_2^*) as a function of the problem parameters, T, θ, ζ_1, r_1, r_2, and N. However, to do so, we make

use of a Taylor expansion of the exponential, which is only valid for small values of the expression $\theta\zeta_1 T n_1(r_1+r_2) - \theta\zeta_1 T r_2 N$.

Given that $0 \leq n_1 \leq N$, for small $\epsilon > 0$ we shall require, for the expansion to hold, that:

$$-\epsilon \leq \theta\zeta_1 T n_1(r_1+r_2) - \theta\zeta_1 T r_2 N \leq \epsilon.$$

Condition 2 *To apply the Taylor expansion with a given precision ϵ, we shall require that the logit scaling parameter θ satisfies:*

$$\theta \leq \frac{\epsilon}{\zeta_1 T N \max(r_1, r_2)}.$$

Remark 1 *Condition 1 becomes:*

$$e^{-\epsilon}\left(r_2 - r_1 - \epsilon(r_1+r_2) + r_2 e^{-\epsilon}\right) \leq r_1 \leq e^{\epsilon}\left(r_2 - r_1 + \epsilon(r_1+r_2) + r_2 e^{\epsilon}\right).$$

Thus, as ϵ tends towards 0, we have that the two price levels must converge to a single price, i.e. $|r_1 - r_2| \leq D(\epsilon)$ with $\lim_{\epsilon\to 0} D(\epsilon) = 0$ and $D'(\epsilon) < 0$. It is not necessary, however, for $\epsilon \to 0$, as it is sufficient that ϵ be small for the expansion to be valid.

The Taylor expansion of the exponential gives:

$$e^{\theta\zeta_1 T n_1(r_1+r_2) - \theta\zeta_1 T r_2 N} = 1 + \theta\zeta_1 T n_1(r_1+r_2) - \theta\zeta_1 T r_2 N + o(n_1^2).$$

The Equation (17) thus becomes:

$$1 + \theta\zeta_1 T n_1(r_1+r_2) - \theta\zeta_1 T r_2 N = \frac{1}{\sqrt{r_2}}\sqrt{\frac{H(n_1)^2}{4r_2} + r_1} - \frac{H(n_1)}{2r_2},$$

as $o(n_1^2)$ is very small. We obtain:

$$1 + \theta\zeta_1 T n_1(r_1+r_2) - \theta\zeta_1 T r_2 N + \frac{H(n_1)}{2r_2} = \frac{1}{\sqrt{r_2}}\sqrt{\frac{H(n_1)^2}{4r_2} + r_1},$$

$$\left(1 + \theta\zeta_1 T n_1(r_1+r_2) - \theta\zeta_1 T r_2 N + \frac{H(n_1)}{2r_2}\right)^2 = \frac{H(n_1)^2}{4r_2^2} + \frac{r_1}{r_2}.$$

Hence we have:

$$(An_1 + B)^2 + (An_1 + B)\frac{H(n_1)}{r_2} = \frac{r_1}{r_2},$$

with $A = \theta\zeta_1 T(r_1+r_2)$ and $B = 1 - \theta\zeta_1 T r_2 N$. By developing the expressions we obtain the following polynomial:

$$n_1^2\left(A^2 + \frac{A\theta\zeta_1 T(r_1+r_2)^2}{r_2}\right)$$

$$+n_1\left(2AB+\frac{A}{r_2}(r_2-r_1-\theta\zeta_1 T r_2 N(r_1+r_2))+\frac{B}{r_2}\theta\zeta_1 T(r_1+r_2)^2)\right)$$

$$+B^2+\frac{B}{r_2}(r_2-r_1-\theta\zeta_1 T(r_1+r_2)r_2 N)-\frac{r_1}{r_2}$$

where the first term is equal to:

$$\theta^2\zeta_1^2 T^2(r_1+r_2)^2(2r_2+r_1).$$

The second term is:

$$2\theta\zeta_1 T(r_1+r_2)(2-2\theta\zeta_1 T r_2 N-\theta\zeta_1 T r_1 N).$$

and the constant term is:

$$2-2\frac{r_1}{r_2}+\theta\zeta_1 T r_2 N(-4+\theta\zeta_1 T r_2 N+\theta\zeta_1 T(r_1+r_2))$$

After some manipulation, we obtain an explicit expression for n_1^*:

$$n_1^*=\frac{(2\theta\zeta_1 T r_2 N+\theta\zeta_1 T r_1 N-2)}{\theta\zeta_1 T(r_1+r_2)(2r_2+r_1)}\pm \frac{\sqrt{\left((2-2\theta\zeta_1 T r_2 N-\theta\zeta_1 T r_1 N)^2-(2r_2+r_1)(2-2\frac{r_1}{r_2}+\theta\zeta_1 T r_2 N(-4+\theta\zeta_1 T r_2 N+\theta\zeta_1 T(r_1+r_2)))\right)}}{\theta\zeta_1 T(r_1+r_2)(2r_2+r_1)}.$$

We may now compare this analytic solution to the Taylor expansion approximation with the exact solution which we obtained numerically. Consider an example with $\theta = 0.05$, $\zeta_1 = 1$, $\zeta_2 = 2$, $r_1 = 2$, $r_2 = 3$, $N = 1$ and $T = 2$. We obtain $n_1^* = 0.1973$ with a revenue of 2.6007 and using the Taylor expansion approximation, we obtain $n_1^* = 0.2131$ with a maximum revenue 2.6004.
Thus, the optimal number of slots for class 1 obtained through the approximation has an error of $\Delta n_1^* = 8\%$, and the revenue difference, taking into account class 2 as well, is essentially, zero.

Figure 4 shows yet a different example, here with three price levels, i.e. $K = 3$. The parameters for this numerical example are $\theta = 0.05$, $\zeta_1 = 1$, $\zeta_2 = 2$, $r_1 = 2$, $r_2 = 4$, $r_3 = 6$, $N = 10$ and the workload is expressed by the following matrix:

$$T=\begin{pmatrix}2 & 4\\ 3 & 4\\ 3 & 5\end{pmatrix}$$

The solution is $n_1^* = 5.2281$, $n_2^* = 2.9909$ and $n_3^* = 1.8110$.

While we observe from Figure 5 that the solution map is neither concave nor quasi-concave (i.e, its level sets are not convex), it is a "nice" non-convex function in that a standard gradient ascent algorithm will generally converge to the global maximum, as we can observe as well from the two figures and from the numerical experience.

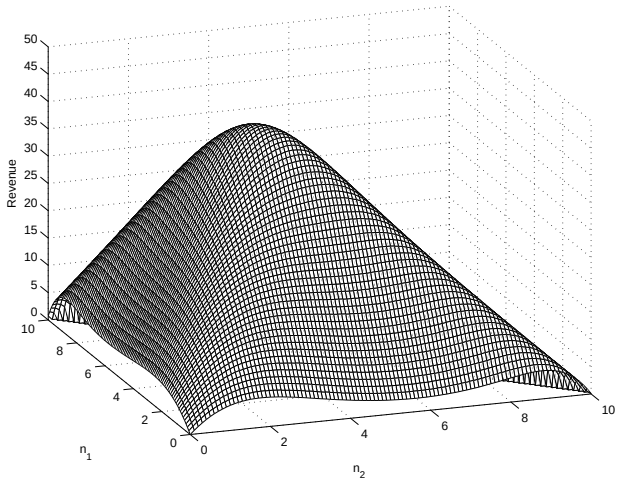


Figure 4: Solution with 3 classes and logit discrete choice model.

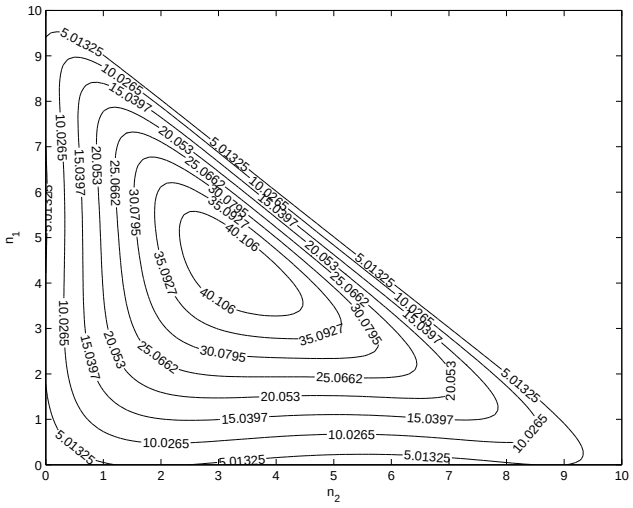


Figure 5: Level sets of the solution with 3 classes and the logit discrete choice model.

4.3 Induced demand curve

In this section, we examine the induced demand function. The demand function is a (decreasing) curve of total demand as a function of price. We illustrate first the induced demand-price curve, in 2-dimensions. The curve is generated by determining the expected quantity that would subscribe at a given price, all other data being fixed. Then, this is repeated at a number of different prices to permit tracing a curve. In Figure 6, the demand curve (solid line) is generated by running the optimization model with the logit choice function. The induced demand curve generated with the deterministic choice function has similar properties. The dotted line above illustrates the revenue curve with price, that is $R(p) = pd(p)$, where $d(p)$ is the induced demand curve shown below it.

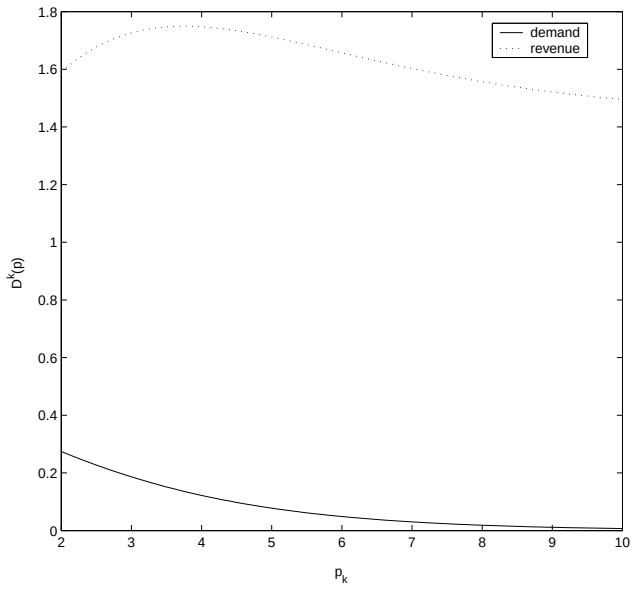


Figure 6: Demand function for the class k with the logit discrete choice model and $\theta = 0.5$.

In Figure 7, we plot the "analytic" optimal yield management solution, which illustrates the increase in revenue as the number of price segments increases, taking into account the induced demand curve. That is, each point on the optimal revenue curve is obtained by plotting the optimal revenue when the given number of price segments is determined optimally (in terms of the actual prices to offer and quantities to offer at each price). These curves, obtained through the use of the induced demand curve, are quite useful in pointing out where the tradeoff in increasing complexity due to a high number of price segments is balanced by a revenue that increases very little.

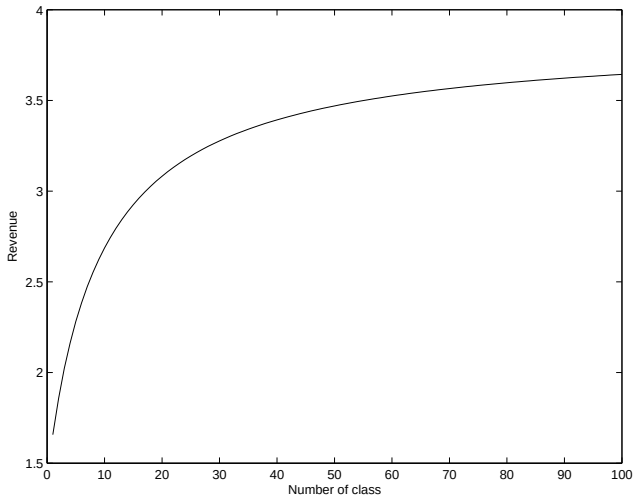


Figure 7: Total revenue as a function of the number of price segments used, where the location and quantities offered at each price segment are optimized.

It is possible to induce such a demand curve in 3-dimensions as well, by taking into account both price and service quality on separate axes. Figure 8 illustrates such a 3-dimensional demand curve which depends on the unit price and on the sojourn time separately.

5 Yield management for web transaction data

We apply our optimization module to web transaction data over an eight-day horizon. Since we are interested here in High Performance Computing (HPC) jobs, rather than short-lived web transactions, we shall suppose that each transaction represents an HPC request. The data we have does not include job durations; therefore we consider all jobs to have unit duration (here, the time unit is one hour). The Yield Management Reservation (YMR) system functions similarly when jobs have heterogeneous durations.
The subscription works as follows: some users, not willing to pay high prices for service, subscribe only if they can obtain the service at an acceptable price level to them. If no such acceptable price is available (not offered, or the maximal quantity is attained) then those customers "go elsewhere". Other users with higher willingness-to-pay can still subscribe, until their threshold is reached, and so on. Therefore, depending upon the prices offered, and the available quantities of each, a different share of the market can be captured, and

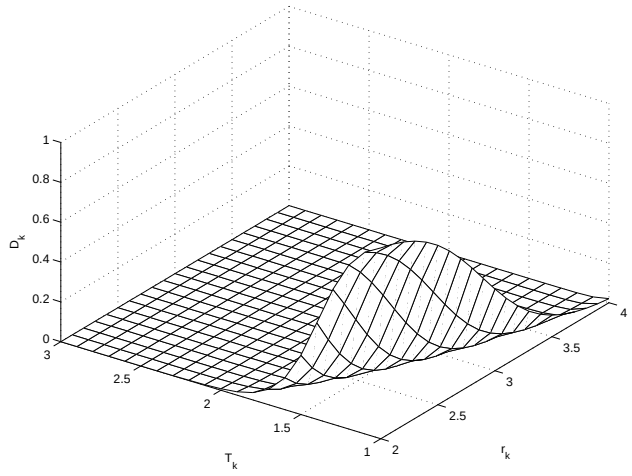


Figure 8: 3-dimensional demand function, in price and sojourn time, for some class k using the logit discrete choice model with $\theta = 0.5$.

revenue will thus vary as well.

The objective of the YMR system is to therefore determine which offerings to propose to customers, and the optimal quantity to propose of each offering, so as to maximize potential revenue. To make an analogy with another existing form of yield management, passenger airline reservation systems make use of similar techniques, whereby they offer limited numbers of seats for each price class on each flight, so as to maximize the airline's expected revenue.

In the next section, we illustrate the output of the YMR system in terms of the optimal number of slots to propose at each of the price levels, and then compare the resulting revenue stream with the base-case, in which a single price per QoS is charged.

The transaction data represents the demand at each point of the time. The YMR model allows for the possibility that a user does not accept any of the offerings proposed. In this series of examples, we have considered a single QoS level and multiple prices for that QoS, with the quantities of slots available at each price limited, by a number to be determined by the YMR. The input data, in addition to the time-varying demands, are as follows. Possible price levels are determined in advance, with not necessarily all price levels open in the optimal solution. On the left column of the table, we consider a variable number of price segments, from 1 single price to 6 price classes; the unit prices themselves are listed to the right in the table 5.

K	Prices
low	.2
medium	.6
high	1
2	.4 .8
3	.3 .6 .9
4	.2 .4 .6 .8
5	.2 .4 .6 .8 1
6	.2 .35 .5 .65 .8 .95

Table 1: Input data on the possible prices for each simulation, in which 1 to 6 price segments are offered to customers, in limited quantities

The first figure illustrates the optimal revenue over time when 2 to 6 price segments are made available to customers, in limited quantities. Possible prices are listed in table 1 above. Note that the topmost curve is the total demand, not the revenue. The revenue accrued under each simulation (2 through 6 price segments on offer) is indicated in the lower series of curves. The larger numbers of price segments (5-6) clearly gives higher revenue during peak periods, whereas during periods of lower use, 2-3 price segments on offer is optimal.

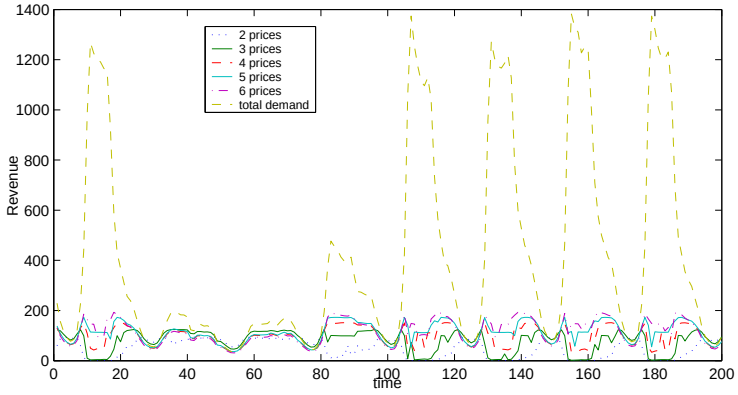


Figure 9: Revenue stream for different numbers of price segments on offer.

The highest curve gives the total demand over time, and serves only to illustrate the peaks and valleys. Higher numbers of segments on offer maximizes revenue when demand is high; for low demand (valleys) a more modest number of price segments is optimal.

Figure 10 summarizes the data of the first two figures for certain time periods, for increased clarity. In particular, we have chosen 5 time periods, with alternating peak flows and off-peak flows, to illustrate how the optimal number of price segments to offer varies.

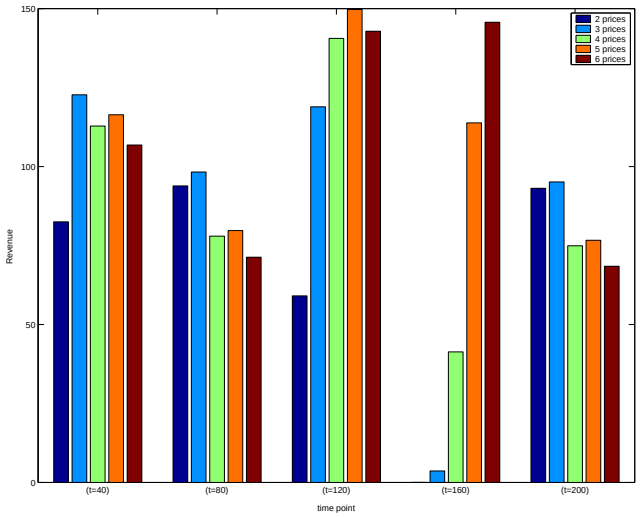


Figure 10: Optimal revenue for 5 different time periods (periods off-peak(40), medium (80), peak (120), peak (160) and off-peak (200)) over the 5 different YMR strategies (offering 2-6 price segments).

Note that higher numbers of segments on offer are optimal for higher-load periods.

The last set of figures compares the revenue when only one price segment is offered (for three cases: a low, medium or high price) with a strategy of offering five classes of price (irrespective of the demand level), shown in the blue-green curve.

Observe that the 5-segment offering is always superior to offering a single price, irrespective of whether a low, medium, or high single price is offered. Furthermore, from the above figures, we know that the YMR system would not suggest always proposing 5 price segments irrespective of the load level, but would allow further revenue increase by modulating the number of segments to offer with the demand level (less segments when demand is low, more when it is high).

Figures 12 illustrates the optimal numbers of slots to offer at each of those 5 price levels and demonstrates that the optimal number of price segments varies with the total demand, or load, in that the higher the demand, the higher the number of segments should be to maximize revenue. This implies furthermore that the YMR should be re-run as new and

better demand data become available. Figure 5 shows the entire breakdown over the 8-day time horizon.

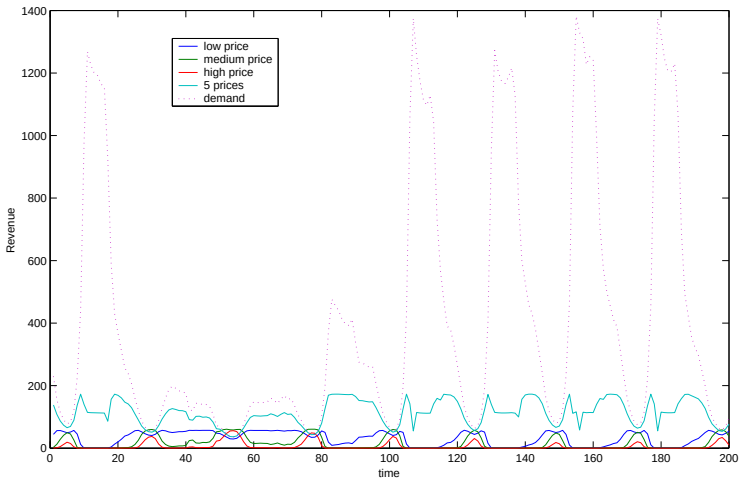


Figure 11: Comparison of YMR strategy of offering 5 price segments with a single-price offering, where the single price is either low, medium, or high.

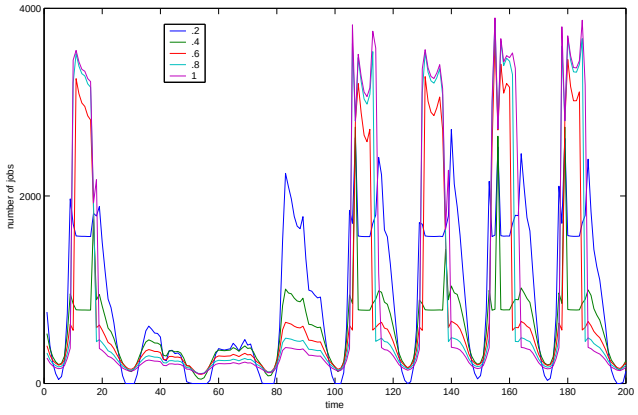


Figure 12: Number of slots to offer at each of the 5 prices segments over time.

6 Future extensions: time incorporated into the model

To incorporate the problem with the time and the redefined constraint. We illustrate the model time by considering only two time periods, i.e. $I = 2$. Hence the objective function is defined by the following:

$$F(n_{11}, n_{12}, n_{21}, n_{22}) = \sum_{i=1}^{2}\sum_{c=1}^{2}\sum_{k=1}^{2} T_{ick} r_{ik} n_{ik} P_{kc} \Gamma_c.$$

Let N_i be the number of free nodes at time i and with $N_1 = N$. Then, we have one constraint for each period, i:

$$N_i - \sum_{k=1}^{K} n_{ik} \geq 0.$$

Furthermore, the number of available nodes at time i is the total number of available nodes N less the nodes allocated for jobs which begun at time z for $z < i$ and whose workload duration is greater than $i - z$. That is, we have the constraint:

$$N_i = N - \sum_{k=1}^{K}\sum_{z=1}^{i-1} n_{zk} \mathbb{1}_{\{\sum_{c=1}^{2} T_{zck} \geq i-z\}}.$$

There are four different cases to consider in our two-period, two-service-class, and two-price-segment example.
If both $(T_{111} + T_{122})$ and $(T_{112} + T_{122})$ are lower than 1, we can apply our previous single-period results, and we obtain the optimal values of the four decision variables n_{ik}, for $i, k = 1, 2$:

$$(n_{11}, n_{12}, n_{21}, n_{22}) = (N\frac{\sqrt{r_{12}}}{\sqrt{r_{11}}+\sqrt{r_{12}}} + \mathcal{H}_1, N\frac{\sqrt{r_{11}}}{\sqrt{r_{11}}+\sqrt{r_{12}}} + \mathcal{H}_1, N\frac{\sqrt{r_{22}}}{\sqrt{r_{21}}+\sqrt{r_{22}}} + \mathcal{H}_2, N\frac{\sqrt{r_{21}}}{\sqrt{r_{21}}+\sqrt{r_{22}}} + \mathcal{H}_2),$$

with $\mathcal{H}_1 = \frac{\zeta_2}{\zeta_1}\frac{\sqrt{r_{12}}}{\sqrt{r_{11}}+\sqrt{r_{12}}}(1 - \frac{\sqrt{r_{12}}}{\sqrt{r_{11}}})$ and $\mathcal{H}_2 = \frac{\zeta_2}{\zeta_1}\frac{\sqrt{r_{22}}}{\sqrt{r_{21}}+\sqrt{r_{22}}}(1 - \frac{\sqrt{r_{22}}}{\sqrt{r_{21}}})$.
The most interesting case is when either T_{11} or T_{12} is greater than 1. Without loss of generality, consider the case where $T_{11} > 1$ and $T_{12} < 1$. We obtain the following system with 6 equations and 6 variables:

$$\begin{cases} (\sqrt{r_{11}}(\zeta_1 r_{12} n_{12} + \zeta_2))^2 - (\sqrt{\frac{\mu_1}{2R_1}}(\zeta_1 r_{11} n_{11} + \zeta_1 r_{12} n_{12} + 2\zeta_2))^2 = 0 \\ (\sqrt{r_{12}}(\zeta_1 r_{11} n_{11} + \zeta_2))^2 - (\sqrt{\frac{\mu_1+\mu_2}{2R_1}}(\zeta_1 r_{11} n_{11} + \zeta_1 r_{12} n_{12} + 2\zeta_2))^2 = 0 \\ (\sqrt{r_{21}}(\zeta_1 r_{22} n_{22} + \zeta_2))^2 - (\sqrt{\frac{\mu_2}{2R_2}}(\zeta_1 r_{21} n_{21} + \zeta_1 r_{22} n_{22} + 2\zeta_2))^2 = 0 \\ (\sqrt{r_{22}}(\zeta_1 r_{21} n_{21} + \zeta_2))^2 - (\sqrt{\frac{\mu_2}{2R_2}}(\zeta_1 r_{21} n_{11} + \zeta_1 r_{22} n_{12} + 2\zeta_2))^2 = 0 \\ n_{11} + n_{12} = N \\ n_{12} + n_{21} + n_{22} = N \end{cases}$$

Note that the first four equations are quadratics; hence each gives two different equations, and we must solve 16 different systems of equations. Generalizing this to a larger system does not, therefore, appear practical.

How often to update segment sizes? The first question to pose when implementing a time-varying version of the yield management optimization model is the granularity at which the decisions are re-optimized. In our setting, this means concretely how long each time period will be in which the number of each type of price segment is held fixed. It may mean that some jobs will wait in a queue if the maximum number of jobs in their highest-cost segment are already present. Figure 13 illustrates the effect of increasing the duration of each time period (homogeneously) on the maximum expected revenue.

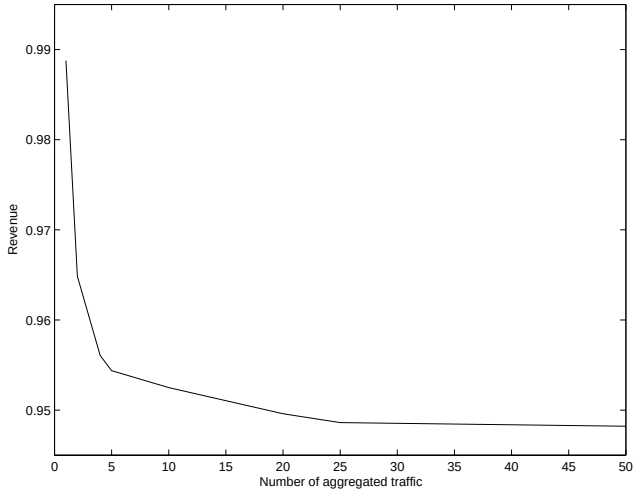


Figure 13: Revenue depending on the aggregation of the traffic

In Figure 14, we look at the dynamic optimization over the time. We compare the case where each epoch has a fixed duration and the case where each epoch has the duration of the maximum workload time. We assume that the duration for each job follows a uniform distribution over the interval $[0, 2]$.

An interesting observation to draw from Figure 14 is that the optimal expected revenue is higher when the duration of each time period is fixed and strictly less than that of the longest job (here set to 2 time units). This is the case due to the distribution of the sojourn times, T, because if the epoch duration is fixed, we can reallocate only the most expensive class. If the duration is equal to the longest job length (2 in this case), all jobs will be treated but it will take too much time. There is a compromise between the duration T and

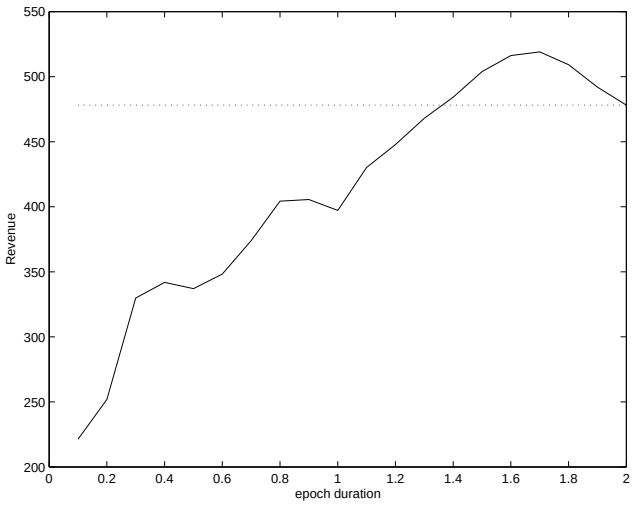


Figure 14: Allowing the duration of a time period to vary, we see the evolution of the optimal expected revenue. The flat, dotted line gives the revenue that one can obtain if the duration of a time period is equal to the longest job length.

the proportion of jobs finishing before next optimization. The revenue is optimized when the proportion of jobs finishing is high and duration time is relatively short.
In Figure 15, we examine the total revenue when we compare a single yield management optimization at the start of the time horizon to more than one, where we let the number of optimizations performed increase to 50 in this figure. Here we assume that the durations of each epoch are not fixed a priori, but depend upon the number of optimizations to be performed over the time horizon.

We remark as the number of optimization increases, the revenue does the same. This property is natural and intuitive, because more and more people are accepted at their optimal price levels, as the demand evolves over time.

7 Conclusion

In this paper, we have applied a yield management model for *On Demand* IT utilities. We have focused our work on an optimal reservation of resources in order to maximize expected revenue. Our behavior model of the user demand was based on one of two discrete choice functions: a deterministic discret choice model and the logit choice model. We have provided a detailed analytical analysis of the optimization problem in both cases when the number of class of prices is small and done so numerically on larger problems. Finally, we provide a

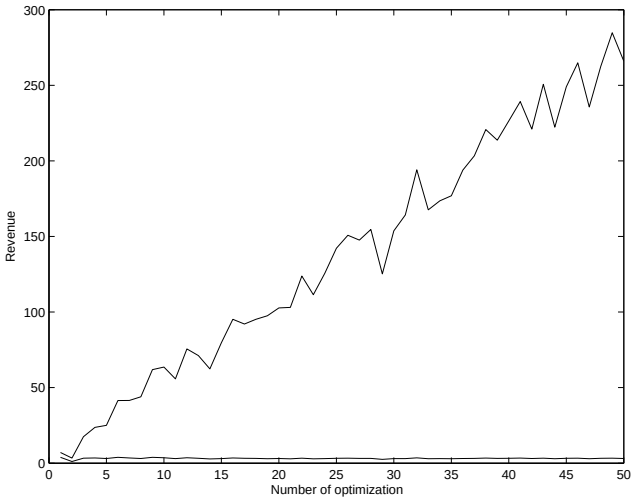


Figure 15: Revenue depending on the number of optimizations. The flat curve at the bottom of the figure is the revenue obtained when the time-zero yield management configuration is used throughout the time horizon.

result on the impact of the number of price segments on the revenue. The model was tested on real web transaction data.

References

[1] P.P. Belobaba, Airline yield management: an overview of seat inventory control, *Transportation Science*, 21, 63–73, 1987.

[2] M. Ben Akiva and S. Lerman *Discrete choice Analysis: Theory and application to travel demand,* MIT Press, Cambridge, 1985.

[3] P. Davis, Airline ties profitability yield to management, *SIAM News*, 27-5, 1994.

[4] R. El Azouzi, A. Altman, and L. Wynter, Telecommunications network equilibrium with price and quality-of-service charactersitics, to appear in the *Proceedings of the 18th International Teletraffic Conference (ITC)*, 2003.

[5] S.E. Kimes, D.I. Barrash and J.E. Alexander, Developing a restaurant revenue management strategy, *Cornell Hotel and Restaurant Administration Quarterly*, 40-5, 18–29, 1999.

[6] S.E. Kimes, Revenue management on the links: applying yield management to the golf course, *Cornell Hotel and Restaurant Administration Quarterly*, 41-1, 120-127, 2001.

[7] A.J. Kleywegt, An optimal control problem of dynamic procing, Georgia Tech Research Report, 2001.

[8] K. Littlewood, Forecasting and control of passengers, 12th AGIFORS Symposium Proceedings, 95–128, 1972.

[9] Z. Liu, L. Wynter, and C. Xia, "Pricing information services in a competitive market: avoiding price wars", INRIA Research Report 4679. Available at www.inria.fr/rrrt/rr-4679.html. Also in proc. of 4th ACM conference on Electronic commerce, San Diego, CA, USA, June 2003

[10] J. Mackie-Mason and H. Varian, Pricing the Internet, in *Public Access to the Internet*, B. Kahn and J. Keller, (Eds.) Prentice Hall, Englewood Cliffs, 1995.

[11] G. van Ryzin and G. Vulcano, Simulation-based optimization of virtual nesting controls for network revenue management, Columbia Business School Working Paper DRO-2003-01, 2003.

[12] K. Talluri and G. van Ryzin, Revenue management under a general discrete choice model of consumer behavior, Columbia Business School Working Paper DRO-2001-02, 2001, to appear in *Management Science*.

[13] L. Wynter, Z. Liu, and P. Dube, "Yield Management for On Demand Computing Services", submitted.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399